Determining sample size for a completely randomized experiment to achieve an acceptable power with a Shiny app

Last edit: December 17, 2024

**Summary** A Shiny app facilitates computing power visualizing the relationship among means, population standard deviation, sample sizes, and significance level. The Shiny app allows students to solve various scenarios by entering these parameters’ and moving the significance slider.

**Keywords:** Power, Shiny app, significance level

# INTRODUCTION

Following the GAISE guidelines (Carver et al., 2016), many instructors have added assessments in the form of small projects to the end of their courses to improve and evaluate student learning. It has been our observation that students will often select one of the last topics they have covered in their course for a project, often comparing many means when given the opportunity. When comparing many means, it is important the study allocate sufficient subjects/experimental units to each treatment to be able to detect differences in treatment means when those differences exist (power). The Shiny package (Chang et al., 2024) can be used to create web based applications. The concept of power and its relationship to sample size for a single sample via a Shiny app was discussed in Arnholt (2019). This work shows how the noncentrality parameter () is a measure of statistical difference between **population** means and how the value computed in an ANOVA table is a measure of statistical difference between **sample** means. A Shiny app is presented where students can experiment with different design structures (i.e. different sample sizes for each of the treatments) to ensure their experiments attain satisfactory power. The notation used for the one-way completely randomized design follows that presented in Ugarte, Militino, & Arnholt (2015). To aid with the connection between noncentrality parameters and test statistics, it is shown how the pooled variance -test is a special case of the -test when there are () treatments before generalizing to treatments.

# STATISTICAL BACKGROUND

The observations in a completely randomized design (CRD) can be described with a linear statistical model

where is the observation of the treatment, is a parameter common to all treatments called the overall mean, is a parameter unique to the treatment called the treatment effect, and is a random error component. For hypothesis testing, the model errors are assumed to be normally and independently distributed with mean zero and constant standard deviation . Although estimating the parameters for Model (1) is possible, the goal of the experimenter is typically to discern whether or not the treatment means are equal. The hypotheses of interest are

The notation that follows is adopted from Ugarte et al. (2015). The sum of the observations in the treatment group is , and the mean of the observations in the treatment group is . The bar indicates a mean while the dot indicates that values have been added over the indicated subscript. The sum of all observations is The grand mean of all observations is denoted For treatments, the typical sum of squares for testing the hypotheses in (2) are shown to the left of the equivalence () symbol in (3). The representation on the right side of the in (3) is the standard test statistic used to test two means when variances are assumed to be equal. To verify the equivalence, keep in mind that and .

Rewriting the right side of (3) yields

The quantity in (3) measures the statistical differences between **sample** means. Let be the sum of squares for treatments obtained by replacing with , with , and with . Then, replacing the sample means in (3) with the population means yields

# EXAMPLE 1

The following scenario can be given to students:

Consider a ficticious experiment where one of two hormones (testosterone or isoandrostenolone) is administered for fifteen days to male chicks starting when the chicks are one day old. At the end of the fifteen days, the experimenter hypothesizes the average weight of chicks that receive testosterone will be 100 mg and that the average weight of chicks that receive isoandrostenolone will be 70 mg. Based on previous work, the researcher estimates the standard deviation for both groups of chicks to be somewhere between 20 mg and 30 mg. What is the minimum number of chicks that should be assigned to each group to obtain a power for the test of at least 0.80 using ?

## Solution using base R functions

Although R (R Core Team, 2024) is used as the computational engine in what follows, most modern software packages can handle what is shown in the code. Start by asking the students to specify the null and alternative hypotheses they will use to test the two hormones. The hypotheses are:

Since the researcher specified a range for , two sets of values will be computed one for when the value of is 20 mg and one for when the value of is 30 mg. Start by finding a critical value for which the null hypothesis will be rejected at the level with a guess of . That is we are looking for . Next compute the noncentrality parameter when mg.

(CriticalF <- qf(0.95, 2 - 1, 20 - 2))

[1] 4.413873

# Power for lambda  
(pf(CriticalF, 2 - 1, 20 - 2, 11.25, lower = FALSE) -> POWER)

[1] 0.8869702

Since we only need to achieve a power of 0.80, we can reduce the values for each sample. Next, we consider using and when mg.

(CriticalF <- qf(0.95, 2 - 1, 9 + 8 - 2))

[1] 4.543077

# Power for lambda  
(pf(CriticalF, 2 - 1, 9 + 8 - 2, 9.5625, lower = FALSE) -> POWER)

[1] 0.8236915

The same calculations are performed next under the assumption that mg.

(CriticalF <- qf(0.95, 2 - 1, 20 - 2))

[1] 4.413873

# Power for lambda  
(pf(CriticalF, 2 - 1, 20 - 2, 5, lower = FALSE) -> POWER)

[1] 0.5620066

Since we need to achieve a power of 0.80, we need to increase the values for each sample. Next, we consider using and when mg.

(CriticalF <- qf(0.95, 2 - 1, 17 + 17 - 2))

[1] 4.149097

# Power for lambda  
(pf(CriticalF, 2 - 1, 17 + 17 - 2, 8.5, lower = FALSE) -> POWER)

[1] 0.8070367

If the standard deviation for chick weight is mg, the experimenter needs to have 17 chicks assigned to each group to detect mean differences between groups greater than 80% percent of the time. If the standard deviation for chick weights is as small as mg, the experimenter can assign 9 chicks to the first group and 8 chicks to the second group and still detect mean differences greater than 80% of the time. To save space, we selected the values of and that solved the problem for the different values of . Next we show how to compute the non-centrality parameter using the notion of the using R, but which will work with any statistical software that computes the sum of squares.

# Hypothesized means  
hypmeans <- c(100, 70)  
# Number of treatments  
a <- length(hypmeans)   
n1 = 17  
n2 = 17  
# Total number of experimental units  
N <- n1 + n2   
# Degrees of freedom for error  
dferror <- N - a  
# Sigma value  
sigma <- 30  
# Create n1 values of 100 and n2 values of 70 and store in Y  
Y <- rep(hypmeans, times = c(n1, n2))  
# Create a treatment factor with n1 values of testosterone   
# and n2 values of isoandrostenolone  
Treat <- factor(rep(c("testosterone", "isoandrostenolone"),   
 times = c(n1, n2)))  
# Compute SS for ANOVA  
summary(aov(Y ~ Treat))

Df Sum Sq Mean Sq F value Pr(>F)   
Treat 1 7650 7650 3.673e+31 <2e-16 \*\*\*  
Residuals 32 0 0   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Pull out the SS Treat value and assign to SShyp  
(summary(aov(Y ~ Treat))[[1]][1, 2] -> SShyp)

[1] 7650

# Noncentrality parameter  
(lambda <- SShyp/sigma^2)

[1] 8.5

(CriticalF <- qf(0.95, a - 1, N - a))

[1] 4.149097

# Power for lambda  
(pf(CriticalF, a-1, dferror, lambda, lower = FALSE) -> POWER)

[1] 0.8070367

## Solution using the Shiny app

To find the required samples sizes to test when mg, launch the Shiny app (Chang et al., 2024) found at <https://shinyapp> and enter the values of 100 and 70 separated with a comma in the box, 10 and 10 separated with a comma in the box, the value of 20 in the box, and use the slider to select a significance level of 0.05 as shown in Figure 1. Change the values for and to be as small as possible with a power value of at least 0.80.

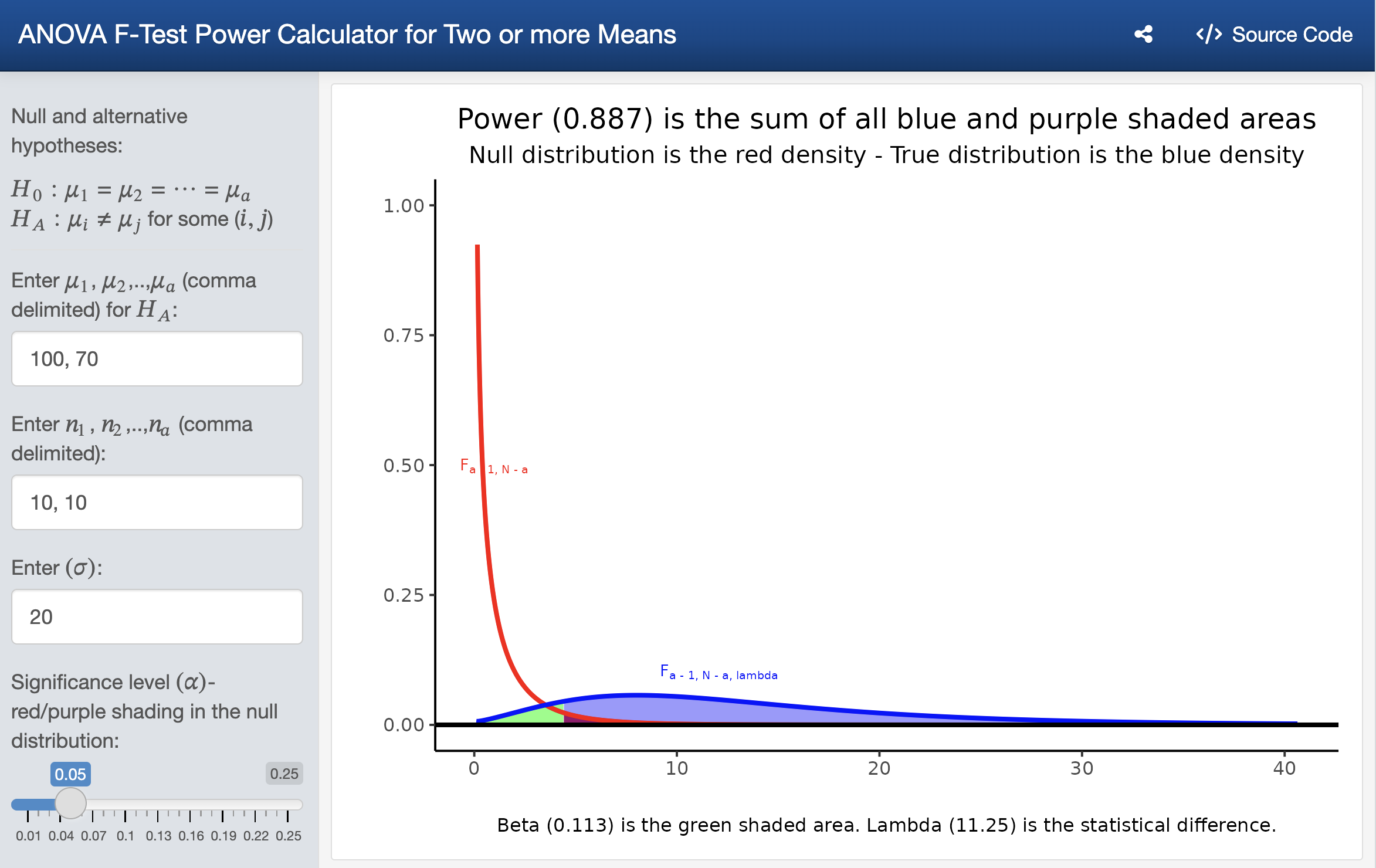


Figure 1: Power to detect a specified difference in means with given sample sizes and population standard deviation

The red density in Figure 1 depicts a central *F*-distribution with and degrees of freedom. The blue density in Figure 1 is a non-central *F*-distribution with and degrees of freedom and non-centrality parameter (. The purple shaded area in Figure 1 is the significance level and the sum of all blue and purple shaded areas is the power (0.887). Changing the values for and to either 8 and 9 or 9 and 8 results in a power of (0.8224) as shown in Figure 2.

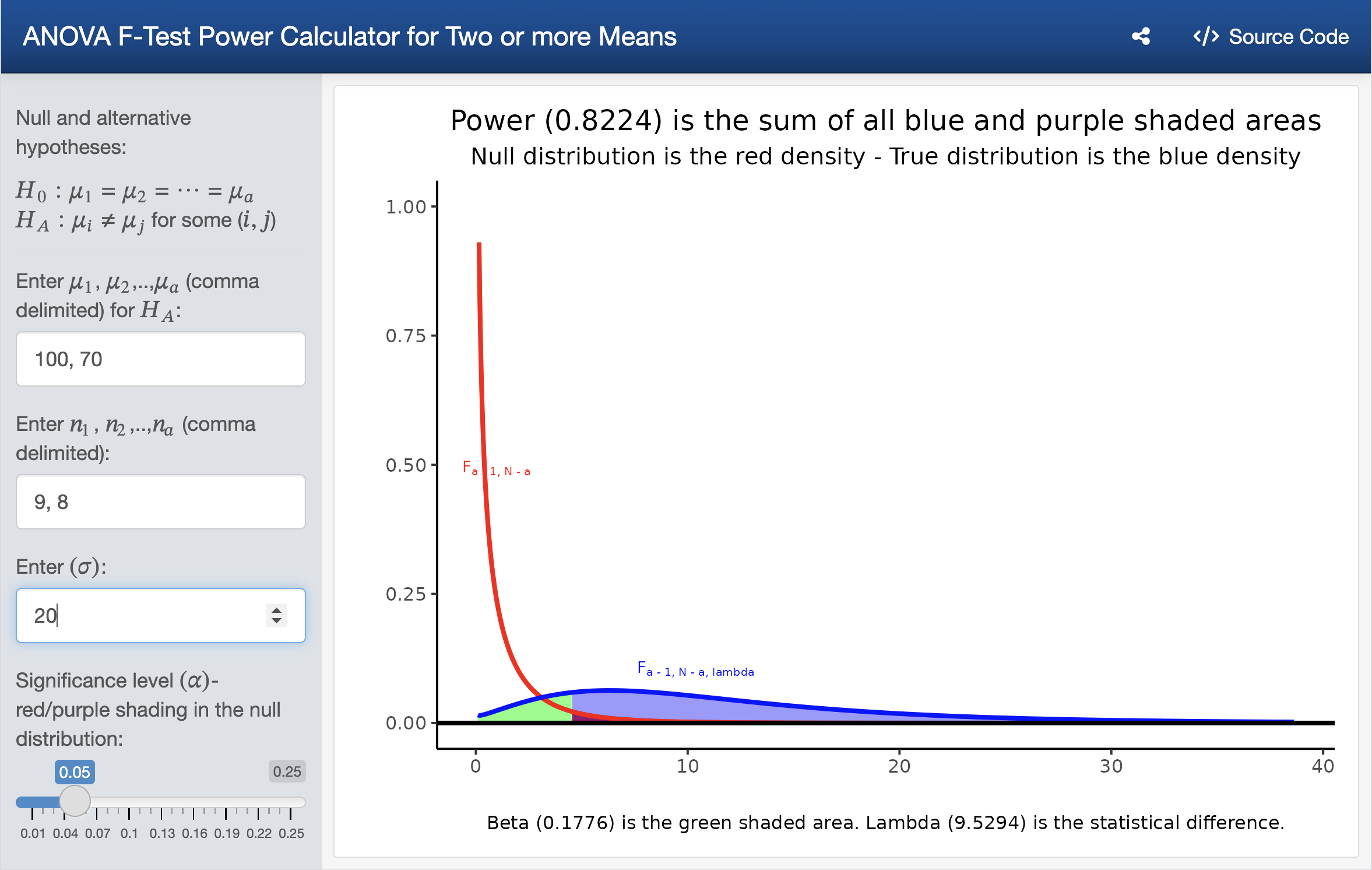


Figure 2: Power to detect a specified difference in means with given sample sizes and population standard deviation

Finally, have the students verify that the minimum samples sizes for and are both 17 when using the Shiny app as shown in Figure 3.

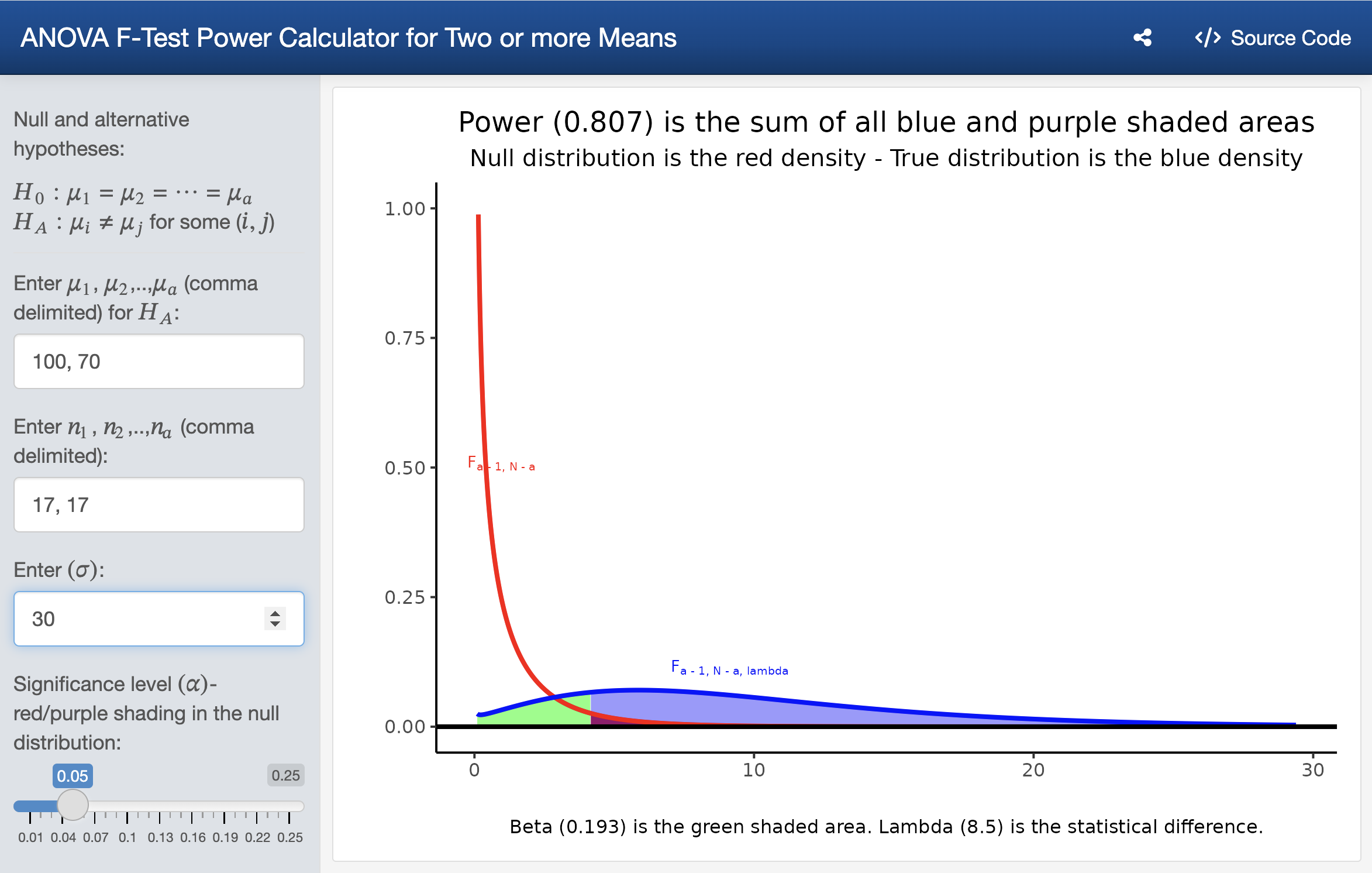


Figure 3: Power to detect a specified difference in means with given sample sizes and population standard deviation

# EXAMPLE 2

An educational researcher is interested in testing different tools to help his students master statistical concepts. The researcher hypothesizes there will be an increasing in mean performance for students on a standardized test from lecture alone (70), to lecture with statistical software (75), to lecture with statistical software and videos (80), to lecture with statistical software, videos, and shiny apps of (85). If the population standard deviation for the researcher’s students on the standardized test is 15, find the minimum number of students required for each group to be able to have a power value of at least 0.80.

Figure 4 shows 70, 75, 80, and 85 entered for , , , and in the box, the value of 15 in the box, and sliding the significance level to 0.05. Power is maximized when all treatment groups have the same number of experimental units. Consequently, if there are no restrictions on the allocation of resources, the experimenter should allocate an equal number of experimental units to each of the treatments. Then, experiment with different values for the sample sizes for each treatment until the power is greater than or equal to 0.80. Our trial and error method started with which yielded too small a power. As a guess, we doubled the values to and noted we were just a little short of our desired power value. Next we tried and observed this was over the target value of 0.80. We continued experimenting lowering the values until we arrived at which returned a power of 0.8082.

Table 1: Process of trial and error for desired power

| n | Power |
| --- | --- |
| 10 | 0.4396 |
| 20 | 0.7856 |
| 25 | 0.8797 |
| 22 | 0.8287 |
| 21 | 0.8082 |

**Answer:** .

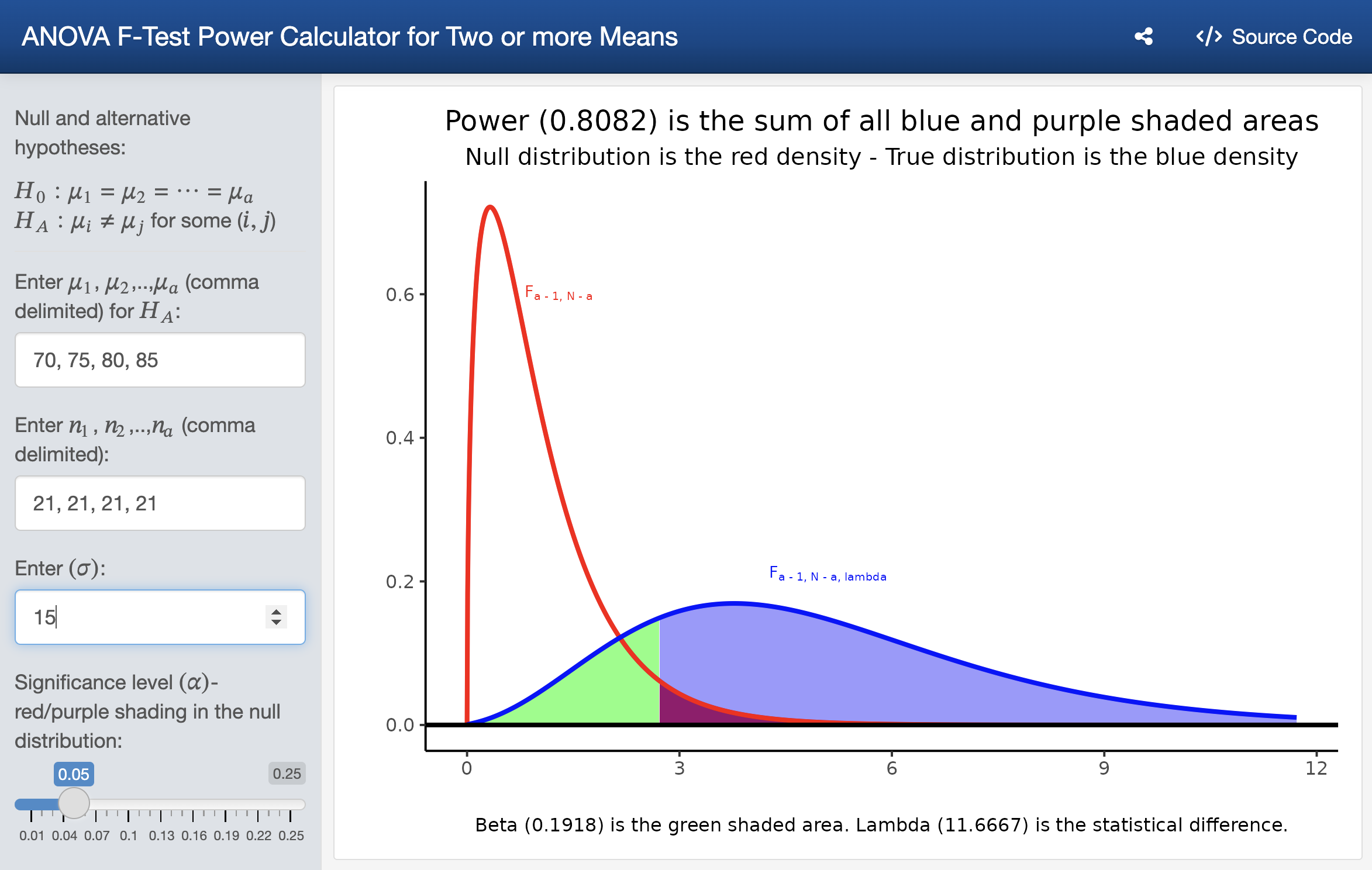


Figure 4: Power to detect a specified difference in means with given sample sizes and population standard deviation

We note in passing that computing power for full-rank general linear models is also possible using the same paradign where . For details see section 12.10 of Ugarte et al. (2015).

# References

Arnholt, A. T. (2019). Using a shiny app to teach the concept of power. *Teaching Statistics*, *41*(3), 79–84. https://doi.org/<https://doi.org/10.1111/test.12186>

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., … Wood, B. (2016). Guidelines for assessment and instruction in statistics education (GAISE) college report 2016.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., … Borges, B. (2024). *Shiny: Web application framework for r*. Retrieved from <https://shiny.posit.co/>

R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>

Ugarte, M. D., Militino, A. F., & Arnholt, A. T. (2015). *Probability and Statistics with R, Second Edition* (2 edition). Boca Raton: Chapman; Hall/CRC.